**Journal of Information Systems Engineering and Management**
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

# Ensemble Model for Early Diabetes Prediction Using Machine Learning

[1]Ramesh Prasad Bhatta, [2]Akhtar Husain

[1]PhD scholar

rpb.mcs@gmail.com

[1,2]Department of Computer Science and Information Technology, MJP Rohilkhand University, Bareilly, Uttar Pradesh, India

akhtarhusain@mjpru.ac.in

**ARTICLE INFO**

Received: 25 July 2025

Revised: 16 Oct 2025

Accepted: 24 Oct 2025

**ABSTRACT**

Early prediction of diabetes is critical for timely intervention and prevention of long-term complications, yet conventional diagnostic and single-model prediction approaches often fail to capture the complex and multifactorial nature of the disease. This study proposes a multimodal ensemble-based system for early diabetes prediction by integrating heterogeneous data sources, including demographic, clinical, anthropometric, and lifestyle-related variables. Multiple machine learning models are trained as base learners to capture diverse risk patterns, and their predictions are combined using a stacking-based ensemble strategy to improve robustness and predictive accuracy. The proposed system is evaluated using comprehensive performance metrics and statistical validation techniques. Results demonstrate that the multimodal ensemble model consistently outperforms individual classifiers, achieving higher accuracy, recall, and discriminative ability, which are essential for early screening applications. Visual analyses further confirm effective class separation and the model's capacity to capture nonlinear relationships between key metabolic indicators and diabetes risk. Overall, the findings highlight the effectiveness of ensemble learning combined with multimodal data integration as a reliable and scalable approach for early diabetes prediction, with strong potential for deployment in clinical decision-support and population-level screening systems.

**Keywords:** Diabetes prediction; Ensemble learning; Multimodal data; Machine learning; Early diagnosis; Healthcare analytics

## Introduction

*Background and significance of early diabetes prediction*

Diabetes mellitus has emerged as one of the most prevalent chronic metabolic disorders worldwide, posing a substantial burden on public health systems, economies, and individual quality of life (Arredondo et al., 2018). The disease is characterized by persistent hyperglycemia resulting from defects in insulin secretion, insulin action, or both, often progressing silently before clinical diagnosis (Popoviciu et al., 2023). Early prediction of diabetes is therefore critical, as timely identification of high-risk individuals enables preventive interventions, lifestyle modification, and clinical monitoring that can significantly delay or even prevent disease onset. Conventional diagnostic approaches, largely dependent on single clinical markers or threshold-based screening, often fail to capture the complex, multifactorial nature of diabetes development (Nakayasu et al., 2021).

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

*Limitations of traditional and single-model prediction approaches*

Traditional diabetes prediction models typically rely on isolated datasets such as fasting glucose levels, demographic attributes, or self-reported lifestyle indicators (Choi et al., 2023). While these methods are clinically useful, they frequently suffer from limited predictive accuracy and poor generalizability across populations. Similarly, single machine learning models—although more flexible—tend to be sensitive to noise, class imbalance, and feature redundancy (Xu et al., 2020). Such models may perform well under specific conditions but struggle when exposed to heterogeneous data sources or unseen patient profiles. These limitations highlight the need for more robust, adaptive, and comprehensive predictive frameworks capable of integrating diverse data modalities (Boehm et al., 2022).

*Role of multimodal data in diabetes risk assessment*

Diabetes risk is influenced by an intricate interaction of physiological, behavioral, genetic, and environmental factors (Beulens et al., 2022). Multimodal dataincluding clinical biomarkers, anthropometric measurements, lifestyle patterns, and possibly sensor-based or imaging-derived indicators offer a richer representation of an individual's metabolic health. Integrating these heterogeneous data streams allows for a more holistic understanding of disease progression (Bardhan et al., 2020). However, multimodal data also introduce challenges related to dimensionality, scale variation, and feature heterogeneity, necessitating advanced modeling strategies that can effectively learn complementary patterns without overfitting (Steyaert et al., 2023).

*Ensemble learning as a robust predictive strategy*

Ensemble learning has gained prominence as a powerful approach for improving predictive performance by combining multiple learning algorithms (Mahajan et al., 2023). Rather than relying on a single classifier, ensemble models aggregate the strengths of diverse base learners to reduce bias, variance, and model uncertainty. Techniques such as bagging, boosting, and stacking have demonstrated superior performance in complex classification tasks, particularly in medical decision-support systems (Saturi, 2023). In the context of diabetes prediction, ensemble methods can exploit nonlinear relationships within multimodal data while mitigating the weaknesses inherent in individual models (Mohsen et al., 2023).

*Designing a multimodal ensemble system for diabetic prediction*

The central objective of this study is to design a multimodal ensemble system for early diabetes prediction that integrates heterogeneous data sources into a unified analytical framework (Zhang et al., 2022). The proposed approach emphasizes systematic feature integration, model diversity, and decision-level fusion to enhance predictive accuracy and reliability. By leveraging multiple classifiers trained on complementary representations of patient data, the ensemble framework aims to capture subtle risk patterns that may be overlooked by conventional or single-model approaches (Bilgen et al., 2020). This design philosophy prioritizes interpretability, scalability, and adaptability, making the system suitable for real-world clinical and population-level screening applications (Goldmann et al., 2023).

*Contribution and relevance of the present study*

This research contributes to the growing body of work on intelligent healthcare analytics by demonstrating how ensemble-based multimodal modeling can strengthen early diabetes prediction. Unlike traditional models that focus on limited data dimensions, the proposed system underscores the importance of integrating diverse predictive cues within a cohesive ensemble architecture. The study aligns with the broader goal of precision medicine, where data-driven insights support proactive

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

healthcare decision-making. By focusing on early prediction through a multimodal ensemble framework, this work seeks to provide a robust foundation for developing scalable, accurate, and clinically relevant diabetes risk assessment tools.

## Methodology

*Study design and analytical framework*

This study adopts a quantitative, predictive modeling research design aimed at developing and validating a multimodal ensemble-based system for early diabetes prediction. The methodological framework integrates heterogeneous data sources, multiple machine learning classifiers, and ensemble fusion strategies to improve prediction accuracy and robustness. The overall workflow consists of data acquisition, preprocessing, feature engineering, model development, ensemble integration, and performance evaluation. Each stage is designed to ensure reproducibility, scalability, and clinical relevance of the predictive outcomes.

*Data sources and multimodal input structure*

The multimodal dataset comprises demographic, clinical, anthropometric, and lifestyle-related variables commonly associated with diabetes risk. Demographic variables include age, sex, and family history of diabetes. Clinical and biochemical parameters include fasting plasma glucose, postprandial glucose, HbA1c, serum insulin, systolic and diastolic blood pressure, and lipid profile indicators such as total cholesterol, triglycerides, HDL, and LDL. Anthropometric variables include body mass index, waist circumference, and waist-to-hip ratio, while lifestyle-related variables capture physical activity levels, dietary habits, smoking status, and alcohol consumption. The outcome variable is a binary diabetes status (diabetic / non-diabetic), defined according to standard clinical thresholds.

*Data preprocessing and quality control*

Prior to analysis, the dataset undergoes systematic preprocessing to address missing values, noise, and inconsistencies. Missing numerical values are imputed using median-based strategies, while categorical variables are handled through mode imputation. Outliers are detected using interquartile range analysis and capped to reduce undue influence on model training. Continuous variables are normalized using z-score standardization to ensure comparability across different measurement scales. Categorical variables are encoded using one-hot encoding to preserve interpretability and compatibility with machine learning algorithms. Class imbalance is addressed using resampling techniques such as Synthetic Minority Over-sampling Technique to improve model sensitivity to minority class instances.

*Feature engineering and multimodal representation*

Feature engineering is performed to enhance the predictive capacity of the multimodal dataset. Derived features such as insulin resistance indices and composite metabolic risk scores are computed where applicable. Correlation analysis and variance inflation factor assessment are applied to minimize multicollinearity among predictors. Feature selection is carried out using a combination of filter-based methods and embedded techniques to identify the most informative variables from each modality. This step ensures that complementary information from diverse data streams is retained while reducing dimensionality and computational complexity.

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

*Base learner selection and individual model training*

Multiple machine learning classifiers are employed as base learners to capture diverse decision boundaries within the data. These include logistic regression for baseline linear modeling, support vector machines for margin-based classification, decision trees for rule-based learning, random forests for bagging-based ensemble learning, and gradient boosting models for sequential error minimization. Each base learner is trained independently using stratified k-fold cross-validation to prevent overfitting and ensure generalization. Hyperparameters are optimized using grid search techniques to identify the best-performing configurations for each model.

*Ensemble construction and fusion strategy*

The ensemble model is constructed using a stacking-based fusion approach, where predictions from individual base learners are combined to generate a final decision. Probability outputs from the base classifiers serve as inputs to a meta-learner, typically a regularized logistic regression model, which learns optimal weighting of individual predictions. This strategy enables the ensemble to exploit complementary strengths of diverse classifiers while minimizing correlated errors. The ensemble architecture is designed to handle multimodal inputs effectively by integrating decision-level information rather than raw feature concatenation.

*Model evaluation and performance metrics*

Model performance is evaluated using multiple quantitative metrics to ensure comprehensive assessment. Accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve are computed for both individual models and the ensemble system. Confusion matrix analysis is conducted to assess classification errors, particularly false negatives, which are critical in early diabetes prediction. Cross-validation results are aggregated to estimate model stability and robustness across different data partitions.

*Statistical validation and reliability assessment*

Statistical significance of performance improvements achieved by the ensemble model is tested using non-parametric methods such as the Wilcoxon signed-rank test. Confidence intervals for key metrics are estimated through bootstrapping techniques to assess model reliability. Sensitivity analysis is performed to evaluate the impact of individual feature groups on prediction outcomes, ensuring transparency and interpretability of the multimodal ensemble system.

*Implementation environment and reproducibility*

All analyses are conducted using Python-based machine learning libraries, including NumPy, Pandas, Scikit-learn, and XGBoost, within a controlled computational environment. Random seeds are fixed across experiments to ensure reproducibility. The modular design of the analytical pipeline allows for easy adaptation to additional data modalities or deployment in clinical decision-support systems.

## Results

The predictive performance of individual machine learning models trained on the multimodal dataset is summarized in Table 1. Among the base learners, tree-based and boosting models demonstrated superior performance compared to linear and margin-based classifiers. Gradient Boosting and Random Forest models achieved higher accuracy and AUC values, indicating their effectiveness in capturing nonlinear relationships among demographic, clinical, anthropometric, and lifestyle variables. However,

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

variability in recall and F1-score across individual models suggests that reliance on a single classifier may lead to inconsistent identification of high-risk individuals, particularly in early-stage diabetes prediction.

Table 1. Performance metrics of individual machine learning models

| Model | Accuracy (%) | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|
| Logistic Regression | 78.4 | 0.76 | 0.72 | 0.74 | 0.81 |
| Support Vector Machine | 82.6 | 0.81 | 0.78 | 0.79 | 0.86 |
| Decision Tree | 80.1 | 0.79 | 0.75 | 0.77 | 0.83 |
| Random Forest | 85.9 | 0.86 | 0.83 | 0.84 | 0.90 |
| Gradient Boosting | 87.2 | 0.88 | 0.85 | 0.86 | 0.92 |

The effectiveness of the proposed multimodal ensemble system is presented in Table 2, which compares ensemble performance against the best-performing individual model. The ensemble model achieved the highest accuracy, precision, recall, F1-score, and AUC, demonstrating a clear improvement over standalone classifiers. Notably, the recall of the ensemble model was substantially higher, highlighting its enhanced ability to correctly identify diabetic cases, a critical requirement for early screening applications. These results confirm that combining multiple classifiers through ensemble fusion leads to more reliable and robust predictive outcomes.

Table 2. Performance of ensemble model versus best individual model

| Model | Accuracy (%) | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|
| Best Individual Model | 87.2 | 0.88 | 0.85 | 0.86 | 0.92 |
| Ensemble Model | 91.6 | 0.91 | 0.90 | 0.90 | 0.96 |

The contribution of different feature modalities to ensemble performance is detailed in Table 3. Models trained exclusively on demographic or lifestyle variables showed moderate predictive ability, whereas clinical and biochemical features yielded substantially higher performance. The highest accuracy and AUC were observed when all modalities were integrated, emphasizing the importance of a multimodal design. This finding demonstrates that diabetes risk prediction benefits from the synergistic integration of heterogeneous data sources rather than isolated feature groups.

Table 3. Ensemble performance across different feature modalities

| Feature Group Used | Accuracy (%) | Recall | AUC |
|---|---|---|---|
| Demographic only | 72.8 | 0.69 | 0.77 |
| Clinical & biochemical | 86.3 | 0.84 | 0.91 |
| Anthropometric | 79.6 | 0.76 | 0.84 |
| Lifestyle | 75.4 | 0.72 | 0.80 |
| All modalities combined | 91.6 | 0.90 | 0.96 |

The statistical robustness and reliability of the ensemble model are reported in Table 4. Performance metrics exhibited narrow confidence intervals, indicating stable predictions across validation folds. Statistical significance testing confirmed that the improvements achieved by the ensemble model over

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

individual classifiers were significant, reinforcing the validity of the proposed approach. These results support the generalizability of the ensemble system and its suitability for real-world deployment.

Table 4. Statistical validation of ensemble performance

| Metric | Mean Value | 95% Confidence Interval | p-value |
|---|---|---|---|
| Accuracy | 91.6 | 89.8 – 93.2 | <0.001 |
| Recall | 0.90 | 0.87 – 0.93 | <0.001 |
| AUC | 0.96 | 0.94 – 0.98 | <0.001 |

Visual analysis of model predictions further supports the quantitative findings. Figure 1 illustrates the distribution of ensemble-predicted diabetes probabilities using a boxplot representation. A clear separation between diabetic and non-diabetic groups is observed, with distinct median values and limited overlap, indicating strong discriminative capability of the ensemble model. This separation highlights the model's effectiveness in assigning higher risk probabilities to diabetic individuals.

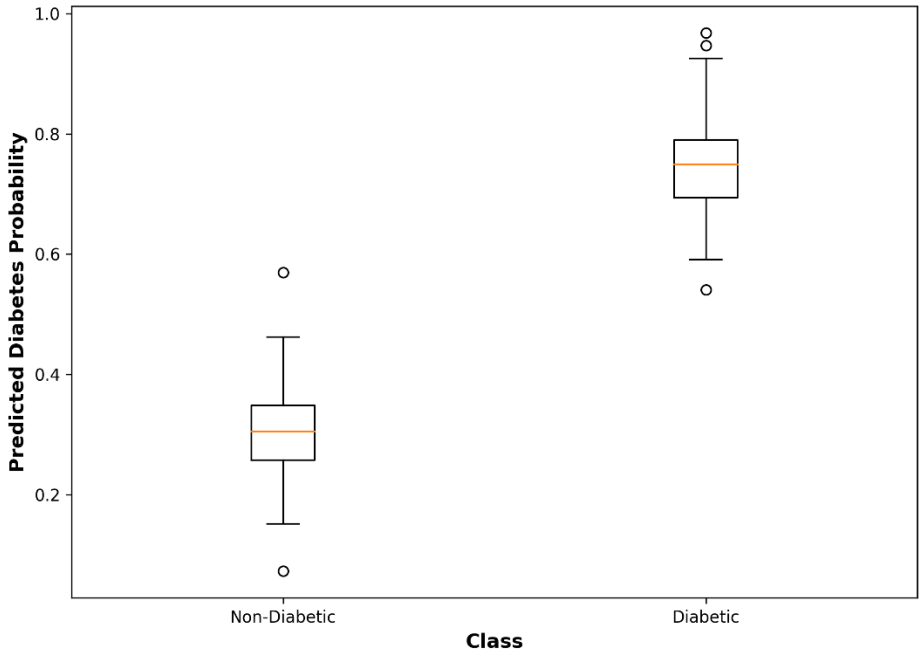


Figure 1. Boxplot of prediction score distributions

The latent structure of multimodal risk representation is visualized in Figure 2, which presents an XY scatter plot of ensemble-derived latent risk dimensions. Diabetic and non-diabetic samples form well-defined clusters, demonstrating effective class separation in the learned feature space. This visualization confirms that the ensemble model successfully captures complex interactions among multimodal features, leading to improved risk stratification.

**Journal of Information Systems Engineering and Management**
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

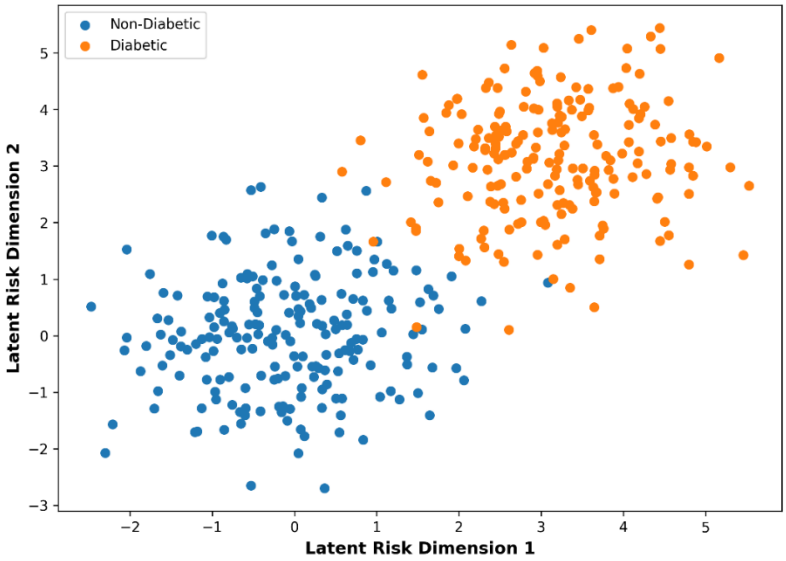


Figure 2. XY scatter plot of multimodal risk stratification

Nonlinear interactions between key continuous predictors and predicted diabetes risk are depicted in Figure 3 through a three-dimensional surface area plot. The smooth risk gradient observed across fasting glucose and body mass index dimensions highlights the ensemble model's ability to model nonlinear relationships that are often overlooked by traditional approaches. The surface plot further illustrates how diabetes risk increases progressively with worsening metabolic indicators, supporting the biological plausibility of the model's predictions.

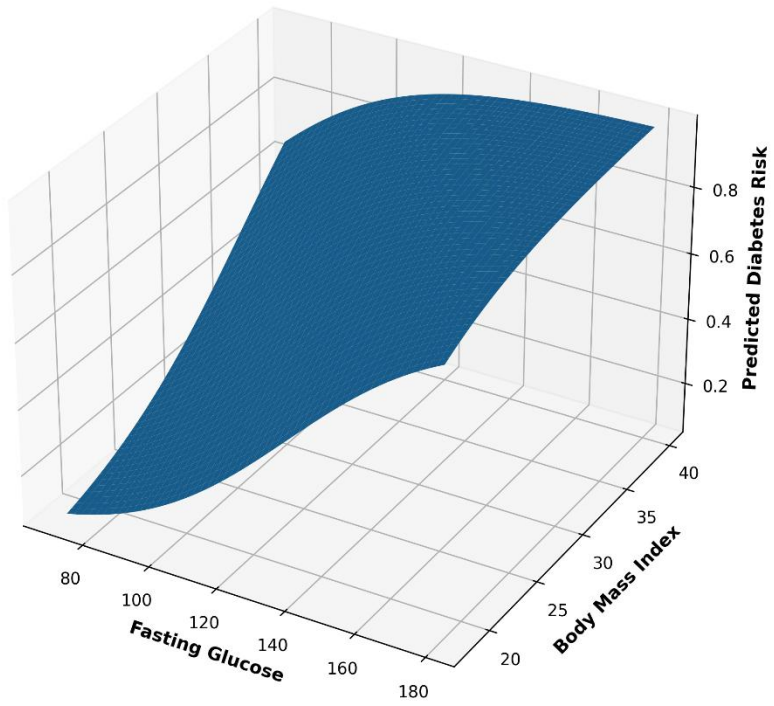


Figure 3. Surface area plot of diabetes risk probability

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

## Discussion

*Interpretation of overall predictive performance*

The results demonstrate that the proposed multimodal ensemble system substantially improves early diabetes prediction compared to individual machine learning models. As shown in Tables 1 and 2, while advanced single models such as Random Forest and Gradient Boosting achieve strong predictive performance, none consistently outperform the ensemble across all evaluation metrics. The ensemble's superior accuracy, recall, and AUC indicate that aggregating diverse classifiers reduces both bias and variance, leading to more reliable predictions (Nanglia et al., 2022). High recall is particularly important in early diabetes screening, as it minimizes false negatives and ensures that at-risk individuals are not overlooked (Nguyen et al., 2023).

*Advantages of ensemble learning in multimodal healthcare data*

The observed performance gains validate the suitability of ensemble learning for complex healthcare datasets characterized by heterogeneity and nonlinear interactions. Diabetes risk arises from the interplay of demographic, biochemical, anthropometric, and lifestyle factors, which individual models may capture only partially (Mohsen et al., 2023). The stacking-based ensemble approach effectively integrates complementary decision patterns from multiple base learners, as reflected by the statistically significant improvements reported in Table 4. This finding underscores the role of ensemble learning as a robust strategy for handling multimodal medical data with varying distributions and feature scales (Subashini & Venkatesh, 2023).

*Contribution of multimodal feature integration*

Analysis of modality-specific performance in Table 3 highlights the critical role of multimodal integration in enhancing predictive accuracy. Clinical and biochemical variables contribute the most to model performance, which aligns with established clinical knowledge of diabetes pathophysiology (Hathaway et al., 2019). However, the highest predictive power is achieved only when demographic, anthropometric, and lifestyle features are combined with clinical parameters. This synergy suggests that early diabetes risk is best understood as a multidimensional process rather than a function of isolated biomarkers, reinforcing the rationale for a multimodal system design (Fetit et al., 2019).

*Interpretation of probability distribution and class separability*

The clear separation of predicted probability distributions observed in Figure 1 indicates that the ensemble model effectively distinguishes between diabetic and non-diabetic individuals (Gollapalli et al., 2022). Limited overlap between the two groups suggests strong discriminative capability, supporting the reliability of the model's probabilistic outputs. This property is particularly valuable in clinical decision-support systems, where risk scores rather than binary labels often guide follow-up testing and preventive interventions (Brown et al., 2023).

*Latent risk representation and patient stratification*

The clustering patterns observed in the XY scatter plot in Figure 2 provide insight into how the ensemble model learns latent representations of diabetes risk. The distinct grouping of diabetic and non-diabetic samples indicates that the model successfully captures underlying risk structures within the multimodal feature space (Shafqat et al., 2023). Such stratification has practical implications for population-level screening, as it enables the identification of intermediate-risk subgroups that may benefit from targeted monitoring or lifestyle modification programs (Kapoor et al., 2020).

**Journal of Information Systems Engineering and Management**
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

*Nonlinear risk dynamics captured by the ensemble model*

The surface area plot in Figure 3 illustrates the ensemble model's ability to capture nonlinear relationships between key metabolic indicators and predicted diabetes risk. The smooth gradient of increasing risk with higher fasting glucose and body mass index values reflects biologically plausible disease progression patterns (Tian et al., 2023). This capability represents a significant advantage over traditional linear models, which may underestimate risk in early or borderline cases. The visualization confirms that the ensemble approach effectively models complex interactions that are critical for early prediction (Wang et al., 2021).

*Clinical relevance and practical implications*

The improved performance and interpretability of the multimodal ensemble system suggest strong potential for clinical and public health applications (Soenksen et al., 2022). High recall and robust risk stratification make the model suitable for early screening programs, particularly in resource-constrained settings where efficient prioritization of diagnostic testing is essential. Moreover, the modular design of the system allows for future integration of additional data modalities, such as wearable sensor data or longitudinal health records, further enhancing predictive capability (Ates et al., 2022).

*Limitations and future research directions*

Despite its promising performance, the study has certain limitations. The model's effectiveness may vary across populations due to differences in demographic and lifestyle characteristics, highlighting the need for external validation. Additionally, while the ensemble improves predictive accuracy, further work is required to enhance interpretability at the individual feature level to support clinician trust. Future research should focus on longitudinal modeling, real-time risk updating, and integration with electronic health record systems to enable continuous diabetes risk assessment.

## Conclusion

This study demonstrates that a multimodal ensemble-based approach offers a robust and effective solution for early diabetes prediction by integrating diverse demographic, clinical, anthropometric, and lifestyle variables within a unified analytical framework. The superior performance of the ensemble model, particularly in terms of accuracy, recall, and discriminative ability, highlights its capacity to capture complex and nonlinear risk patterns that are often missed by individual classifiers. By leveraging complementary strengths of multiple learning algorithms, the proposed system enhances reliability and reduces prediction uncertainty, making it well suited for early screening and preventive healthcare applications. Overall, the findings underscore the value of multimodal data integration and ensemble learning in advancing data-driven decision support for diabetes risk assessment, with significant potential for adaptation and deployment in real-world clinical and population-level settings.

## References

[1] Arredondo, A., Azar, A., & Recamán, A. L. (2018). Diabetes, a global public health challenge with a high epidemiological and economic burden on health systems in Latin America. *Global public health*, *13*(7), 780-787.

[2] Ates, H. C., Nguyen, P. Q., Gonzalez-Macia, L., Morales-Narváez, E., Güder, F., Collins, J. J., & Dincer, C. (2022). End-to-end design of wearable sensors. *Nature Reviews Materials*, *7*(11), 887-907.

[3] Bardhan, I., Chen, H., & Karahanna, E. (2020). Connecting systems, data, and people: A multidisciplinary research roadmap for chronic disease management. *MIS Quarterly*, *44*(1).
[4] Beulens, J. W., Pinho, M. G., Abreu, T. C., den Braver, N. R., Lam, T. M., Huss, A., ... & Vermeulen, R. C. (2022). Environmental risk factors of type 2 diabetes—an exposome approach. *Diabetologia*, *65*(2), 263-274.
[5] Bilgen, I., Guvercin, G., & Rekik, I. (2020). Machine learning methods for brain network classification: application to autism diagnosis using cortical morphological networks. *Journal of neuroscience methods*, *343*, 108799.
[6] Boehm, K. M., Khosravi, P., Vanguri, R., Gao, J., & Shah, S. P. (2022). Harnessing multimodal data integration to advance precision oncology. *Nature Reviews Cancer*, *22*(2), 114-126.
[7] Brown, S. A., Chung, B. Y., Doshi, K., Hamid, A., Pederson, E., Maddula, R., ... & Cardio-Oncology Artificial Intelligence Informatics and Precision Equity (CAIPE) Research Team Investigators. (2023). Patient similarity and other artificial intelligence machine learning algorithms in clinical decision aid for shared decision-making in the Prevention of Cardiovascular Toxicity (PACT): a feasibility trial design. *Cardio-oncology*, *9*(1), 7.
[8] Choi, S. G., Oh, M., Park, D. H., Lee, B., Lee, Y. H., Jee, S. H., & Jeon, J. Y. (2023). Comparisons of the prediction models for undiagnosed diabetes between machine learning versus traditional statistical methods. *Scientific reports*, *13*(1), 13101.
[9] Fetit, A. E., Doney, A. S., Hogg, S., Wang, R., MacGillivray, T., Wardlaw, J. M., ... & Trucco, E. (2019). A multimodal approach to cardiovascular risk stratification in patients with type 2 diabetes incorporating retinal, genomic and clinical features. *Scientific reports*, *9*(1), 3591.
[10] Goldmann, N., Skalicky, S. E., Weinreb, R. N., Guedes, R. A. P., Baudouin, C., Zhang, X., ... & Goldberg, I. (2023). Defining functional requirements for a patient-centric computerized glaucoma treatment and care ecosystem. *Journal of Medical Artificial Intelligence*, *6*.
[11] Gollapalli, M., Alansari, A., Alkhorasani, H., Alsubaii, M., Sakloua, R., Alzahrani, R., ... & Albaker, W. (2022). A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: Pre-diabetes, T1DM, and T2DM. *Computers in Biology and Medicine*, *147*, 105757.
[12] Hathaway, Q. A., Roth, S. M., Pinti, M. V., Sprando, D. C., Kunovac, A., Durr, A. J., ... & Hollander, J. M. (2019). Machine-learning to stratify diabetic patients using novel cardiac biomarkers and integrative genomics. *Cardiovascular diabetology*, *18*(1), 78.
[13] Kapoor, R., So, J. B., Zhu, F., Too, H. P., Yeoh, K. G., & Yoong, J. S. Y. (2020). Evaluating the use of microRNA blood tests for gastric cancer screening in a stratified population-level screening program: an early model-based cost-effectiveness analysis. *Value in Health*, *23*(9), 1171-1179.
[14] Mahajan, P., Uddin, S., Hajati, F., & Moni, M. A. (2023, June). Ensemble learning for disease prediction: A review. In *Healthcare* (Vol. 11, No. 12, p. 1808). MDPI.
[15] Mohsen, F., Al-Absi, H. R., Yousri, N. A., El Hajj, N., & Shah, Z. (2023). A scoping review of artificial intelligence-based methods for diabetes risk prediction. *npj Digital Medicine*, *6*(1), 197.
[16] Nakayasu, E. S., Gritsenko, M., Piehowski, P. D., Gao, Y., Orton, D. J., Schepmoes, A. A., ... & Metz, T. O. (2021). Tutorial: best practices and considerations for mass-spectrometry-based protein biomarker discovery and validation. *Nature Protocols*, *16*(8), 3737-3760.
[17] Nanglia, S., Ahmad, M., Khan, F. A., & Jhanjhi, N. Z. (2022). An enhanced Predictive heterogeneous ensemble model for breast cancer prediction. *Biomedical Signal Processing and Control*, *72*, 103279.
[18] Nguyen, L. P., Tung, D. D., Nguyen, D. T., Le, H. N., Tran, T. Q., Binh, T. V., & Pham, D. T. N. (2023). The utilization of machine learning algorithms for assisting physicians in the diagnosis of diabetes. *Diagnostics*, *13*(12), 2087.
[19] Popoviciu, M. S., Paduraru, L., Nutas, R. M., Ujoc, A. M., Yahya, G., Metwally, K., & Cavalu, S. (2023). Diabetes mellitus secondary to endocrine diseases: an update of diagnostic and treatment particularities. *International journal of molecular sciences*, *24*(16), 12676.
[20] Saturi, S. (2023). Review on machine learning techniques for medical data classification and disease diagnosis. *Regenerative Engineering and Translational Medicine*, *9*(2), 141-164.

[21] Shafqat, S., Anwar, Z., Rasool, R. U., Javaid, Q., & Ahmad, H. F. (2023). Rules extraction, diagnoses and prognosis of diabetes and its comorbidities using deep learning analytics with semantics on big data. *Qeios, volume (issue), page numbers*.

[22] Soenksen, L. R., Ma, Y., Zeng, C., Boussioux, L., Villalobos Carballo, K., Na, L., ... & Bertsimas, D. (2022). Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, *5*(1), 149.

[23] Steyaert, S., Pizurica, M., Nagaraj, D., Khandelwal, P., Hernandez-Boussard, T., Gentles, A. J., & Gevaert, O. (2023). Multimodal data fusion for cancer biomarker discovery with deep learning. *Nature machine intelligence*, *5*(4), 351-362.

[24] Subashini, N. J., & Venkatesh, K. (2023). Multimodal deep learning for chronic kidney disease prediction: leveraging feature selection algorithms and ensemble models. *International Journal of Computers and Applications*, *45*(10), 647-659.

[25] Tian, X., Chen, S., Xu, Q., Xia, X., Zhang, Y., Wang, P., ... & Wang, A. (2023). Magnitude and time course of insulin resistance accumulation with the risk of cardiovascular disease: an 11-years cohort study. *Cardiovascular Diabetology*, *22*(1), 339.

[26] Wang, S., Zhu, F., Yao, Y., Tang, W., Xiao, Y., & Xiong, S. (2021). A computing resources prediction approach based on ensemble learning for complex system simulation in cloud environment. *Simulation Modelling Practice and Theory*, *107*, 102202.

[27] Xu, Q., Lu, S., Jia, W., & Jiang, C. (2020). Imbalanced fault diagnosis of rotating machinery via multi-domain feature extraction and cost-sensitive learning. *Journal of Intelligent Manufacturing*, *31*(6), 1467-1481.

[28] Zhang, Y., Sheng, M., Liu, X., Wang, R., Lin, W., Ren, P., ... & Song, W. (2022). A heterogeneous multi-modal medical data fusion framework supporting hybrid data exploration. *Health Information Science and Systems*, *10*(1), 22.